Jennie B. Chandra
Senior Counsel, Federal Policy
Windstream Communications, Inc.
1101 17th Street, NW, Suite 802
Washington, DC 20036

(202) 223-7667
jennie.b.chandra@windstream.com

June 30, 2011

Marlene H. Dortch
Secretary
Federal Communications Commission
445 12th Street, SW
Washington, DC 20554

Re: *Connect America Fund*, WC Docket No. 10-90; *A National Broadband Plan for Our Future*, GN Docket No. 09-51; *Federal-State Joint Board on Universal Service*, CC Docket No. 96-45; *High-Cost Universal Service Support*, WC Docket No. 05-337; *Lifeline and Link-Up*, WC Docket No. 03-109; *Developing an Unified Intercarrier Compensation Regime*, CC Docket No. 01-92; *Establishing Just and Reasonable Rates for Local Exchange Carriers*, WC Docket No. 07-135

Dear Ms. Dortch:

On June 28, 2011, Jeff Lanning and Melissa Newman of CenturyLink; Jim Stegeman of CostQuest; Mike Saperstein of Frontier; Malena Barzilai and Eric Einhorn of Windstream; and I met with Sharon Gillett, Carol Mattey, Steve Rosenberg, Rebekah Goodheart, Patrick Halley, Joe Cavender, and Katie King of the Wireline Competition Bureau. Ken Mason of Frontier and Cesar Caballero and Bill Kreutz of Windstream joined the discussion via phone.

The companies reiterated their call for immediate adoption of reforms that would distribute ongoing support within price cap carriers' areas on the basis of cost conditions in individual wire centers, rather than costs averaged across study areas or entire states. To enable immediate implementation of these reforms, the companies explained how a regression analysis could be used to forecast the costs to serve individual wire centers in study areas that have been categorized as "non-rural" (wire centers that were included in the original run of the Hybrid Cost Proxy Model), as well as the costs to serve individual wire centers in price cap carriers' "rural"-categorized study areas. A specific proposal for this regression analysis and details on how it was developed and would operate are included in the attached white paper.

Please feel free to contact us if you require any additional information.

Sincerely,

/s/

Jennie B. Chandra

Attachment

cc: Sharon Gillett
Carol Mattey
Steve Rosenberg
Rebekah Goodheart
Patrick Halley
Joe Cavender
Katie King

DEVELOPMENT OF A REGRESSION-BASED ANALYSIS TO FORECAST HCPM COST OUTPUTS FOR USE IN CALCULATING AND DISTRIBUTING SUPPORT TO PRICE CAP AREAS

SUMMARY

The Hybrid Cost Proxy Model ("HCPM") produces cost outputs for price cap carriers' "non-rural" study areas, but it has not been applied to these carriers' "rural" study areas to date. To enable comparisons of costs across all price cap areas, this white paper describes how a relatively simple regression equation can be used to forecast HCPM cost outputs. The paper then proposes a method for using the wire center cost forecast to develop unitized cost estimates for use in targeting universal service support on the basis of cost conditions in price cap carriers' individual wire centers.

CostQuest[1] and KDD[2] developed the regression analysis in a process comprised of two stages. The first stage encompassed development of regression equations that would accurately estimate the output of the 2004 HCPM results. In the second stage, a single regression equation was selected, on the basis of the dual goals of removing as much complexity as possible from the statistical equation and its implementation, while still preserving the accuracy of prediction. This development process made use of actual HCPM wire center cost outputs for more than 11,000 wire centers produced as part of 2004 HCPM Synthesis Model processing, as well as a set of "driver variables" derived from U.S. Census data, HCPM input files, wire center boundaries, and demographics from as close in time to the HCPM development period as possible.

This white paper also includes a proposal for how the Federal Communications Commission ("Commission") can use the selected regression equation, in conjunction with updated inputs, to forecast wire center cost for all price cap carriers' wire centers. This effort effectively would offer a forecast of HCPM outputs in light of changes to underlying density and demographic patterns. These wire center cost outputs then can be unitized and compared to varying levels of benchmarks to establish, based on various benchmarks, estimates of overall universal service funding levels as well as funding levels on a wire center basis.

[1] CostQuest Associates, http://www.costquest.com.

[2] KDD Analytics, http://kddanalytics.com.

DATA

HCPM OUTPUT

The HCPM, which was developed in the late 1990s, produces estimates of the "forward looking" cost of providing telecommunication service by wire center (CLLI), which are then used to determine USF support for non-rural study areas under the Commission's High-Cost Model mechanism. The HCPM considers how an efficient telecommunications network would be built starting from scratch, taking as given the actual locations of households and businesses, and engineering and cost assumptions concerning such a build-out. The HCPM produces per switched access line cost estimates for six subcomponents of the Total Cost of telephone service: loop, end-office usage, transport, lineport, local number porting ("LNP"), and signaling. The loop typically accounts for 85 percent of the Total Cost. For the regression effort we multiplied the per line cost by the total switched lines in a wire center to arrive at the total monthly cost for serving the entire wire center. HCPM cost estimates generated by the Commission were released after parties signed restricted use agreements. These cost data cover more than 11,000 wire centers in all 50 states plus the District of Columbia.

The regression analysis considered the entire Commission wire center data set of more than 11,000 wire centers and estimated the total cost of the wire center (versus a per line value). Ultimately only three wire centers were excluded from the sample on the basis of HCPM output. Two wire centers were excluded due to extreme values for Total Cost.[3] A third wire center was excluded because the HCPM results did not contain an estimate for LNP cost. After these three outlier exclusions, the sample of wire centers available for development of the regression analysis was 11,103 wire centers.

REGRESSION DRIVER VARIABLES

A set of potential regression "driver variables" was developed. Variables were either directly extracted from source data, such as wire center boundaries, or were computed via database roll-ups (such as locations in a wire center) or computed geographically (such as CLLIRoadFt). These driver (or explanatory) variables reflect location and geographic characteristics of wire centers and are primarily variations of wire center size; household and business locations; and road feet (e.g., counts, density measures, counts within concentric rings around the wire center mid-point, etc.). A total of 34 such variables were created for 11,078[4] of the 11,106 wire centers for which HCPM cost estimates are available. Table 1 displays the exploratory variables.

[3] For the purposes of the regression development, cost is expressed on a per month basis. Because the cost estimates to be modeled by regression are from an engineering model and not from a survey, we were reluctant to exclude any additional wire centers as potential outliers.

[4] A small number of CLLIs were excluded because CLLI codes did not match closely in name between the independent and dependent data.

Table 1. Exploratory Variables

Potential Driver Variable	Description
CLLI	Wire center boundary name (CLLI code)
ERS	USDA ERS rurality continuum
CLLIArea	Area of geographic object
CLLIRoadft	Tiger Road Feet Allocated into the CLLI boundary
HH	Res locations from HCPM IN file adjusted to estimated 1997 counts
BusinessLocations	Bus locations from HCPM IN file
LocationDensity	(HH+BusinessLocations)/CLLIArea
RoadDensity	CLLIRoadft/CLLIArea
LocationsPerRoad	HH+BusinessLocations/CLLLIRoadft
CoSize	Company Size indicator (S,M,L)
OCN	Operating Company Number (based upon commercial wirecenter boundary product)
SAC	Study Area Code (based upon OCN)
NearestCOft	Nearest Central Office to the Central Office in this boundary (distance)
CLLIRoadft_18kft	The amount of road feet within 18kft (airline) from Central Office
HH_18kft	The number of households within 18 kft (airline) from Central Office
BusinessLocations_18kft	The number of business locations within 18kft (airline) from Central Office.
LocationDensity_18kft	The number of locations (business locations plus households) within 18kft from Central Office
RoadDensity_18kft	The road density as calculated for Census blocks within 18kft
LocationsPerRoad_18kft	The locations per road as calculated for Census blocks within 18kft
CLLIRoadft_36kft	The amount of road feet beyond 18kft and within 36kft (airline) from Central Office

HH_36kft	The number of households beyond 18kft and within 36 kft (airline) from Central Office
BusinessLocations_36kft	The number of business locations beyond 18kft and within 36kft (airline) from Central Office.
LocationDensity_36kft	The number of locations (business locations plus households) beyond 18kft and within 36kft from Central Office
RoadDensity_36kft	The road density as calculated for Census blocks within beyond 18kft and 36kft
LocationsPerRoad_36kft	The locations per road as calculated for Census blocks within beyond 18kft and 36kft
CLLIRoadft_54kft	The amount of road feet beyond 36kft and within 54kft (airline) from Central Office
HH_54kft	The number of households beyond 36kft and within 54 kft (airline) from Central Office
BusinessLocations_54kft	The number of business locations beyond 36kft and within 54kft (airline) from Central Office.
LocationDensity_54kft	The number of locations (business locations plus households) beyond 36kft and within 54kft from Central Office
RoadDensity_54kft	The road density as calculated for Census blocks beyond 36kft and within 54kft
LocationsPerRoad_54kft	The locations per road as calculated for Census blocks beyond 36kft and within 54kft
CLLIRoadft_Over54kft	The amount of road feet over 54kft (airline) from Central Office
HH_Over54kft	The number of households over 54 kft (airline) from Central Office
BusinessLocations_Over54kft	The number of business locations over 54kft (airline) from Central Office.
LocationDensity_Over54kft	The number of locations (business locations plus households) over 54kft from Central Office
RoadDensity_Over54kft	The road density as calculated for Census blocks over 54kft
LocationsPerRoad_Over54kft	The locations per road as calculated for Census blocks over 54kft

SAMPLE

After merging the wire centers for which HCPM cost estimates are available (excluding the three outliers) with the wire centers for which potential driver variables were created, 11,075 wire centers were available for development of the regression analysis.

REGRESSION METHODOLOGY

SINGLE EQUATION APPROACH

As noted above, the HCPM produces cost estimates for six components of Total Cost. One potential approach, at a high level, would be to estimate a regression equation for each of these six components and sum the predicted costs to yield an estimate of Total Cost.[5] For the following reasons, it was determined that the best approach would be to estimate a single regression for Total Cost of the wire center, and use average actual HCPM shares to estimate the cost for each of the six components:

- First, the set of driver variables available for this effort are more suited to estimating loop cost than many of the other components of Total Cost (e.g. signaling, transport).
- Second, exploratory analysis indicates that the average prediction error rate from regressions estimated for components other than loop (using the available set of driver variables) is higher, sometimes much higher (e.g. 1.6 to 4.0 times higher) than that for loop or Total Cost.
- Third, loop accounts for, on average, 85 percent of the HCPM's estimated Total Cost, and exploratory analysis indicates that a regression for Total Cost yields the lowest average prediction error.
- Fourth, estimating a single regression for Total Cost is consistent with the current Commission method. Component-level costs from the HCPM currently are not used as discrete parts of the support estimate; instead, the components are summed together to yield Total Cost.

DATA TRANSFORMATIONS

In developing the regression analysis, all exploratory variables were expressed as natural logarithms. This approach normalized the variables and minimized the effects on the regression coefficients of highly skewed distributions. Expressing variables in natural logarithms also eliminated the possibility of the regression generating a negative predicted Total Cost.

To expand the pool of potential driver variables, several variables that consistently exhibited strong correlations with Total Cost also were expressed as second orders (i.e., squared) and interacted (multiplied) with each other. Additional variables were created by combining existing variables into one

[5] Such an approach also would require cross-equation restrictions on coefficients, the use of cost shares as dependent variables, and joint estimation methodologies.

(e.g., Total Locations) and expressing existing variables as shares (e.g., business location share of total locations).

MODELING VS. HOLDOUT SAMPLE

To test and validate the predictive capabilities of a potential regression equation, the total sample of 11,075 wire centers was divided into a "modeling" sample and a "holdout" sample. A random sample of 70 percent of the 11,075 (7,762 wire centers) was used for regression development. Regression equation candidates were tested using the remaining 30 percent (3,313 wire centers). This approach ensured that regression equation candidates were not tested (or "validated") on the same data used to estimate the regression equations. Once a regression equation was finalized, it was recalibrated using the entire sample of 11,075 wire centers.

REGRESSION DEVELOPMENT

The regression development process consisted of finding a set of candidate regression equations to test using the holdout sample. These candidate regression equations were developed using both automatic and manual variable selection routines.

Automatic routines – such as stepwise, forward, and backward regression – select variables one at a time for entry into (or removal from) a regression equation based on a user-specified level of statistical significance. Variables selected in an earlier stage are retested to ensure that their level of significance exceeds the user-specified value. Selection continues until all the variables available have been tested and those selected to be in the regression pass the specified significance test. No restrictions are placed on the number of variables that can be selected other than their level of statistical significance (which here was set very high at .0001).

Based on the results of the automatic selection routines, and the contribution of each variable to the regression's overall explanatory power (i.e., "r^2"), additional candidate regression equations can be developed manually. Here the aim was to find regression equations with a smaller number of driver variables that account for at least 95 percent of the automatically selected regression equation's explanatory power. At no point in the process of developing the regression equation or variable selection were results by company analyzed to determine the preferred equation, variable or parameters.

REGRESSION CRITERIA

Because the goal of this effort was to develop a regression equation that can reliably predict the output of the HCPM, candidate regression equations were judged based on their predictive power in the holdout sample. Specifically the regression's cost estimates for wire centers in the holdout sample were compared to actual 2004 HCPM outputs.

OVERALL FINDINGS

Seven different candidate regressions were examined: three using automatic variable selection and four using manual variable selection. A comparison of these regressions determined the following:

- The explanatory power of the seven regressions was very similar – with adjusted r^2 between .87 and .92. The automatic selection regression equations typically produced somewhat higher r^2

values than the manual selection regression equations, due to the higher number of variables in the automatic selection regressions.

- The predictive power of the regression equations also was relatively comparable, with little difference among various regression equations' overall ability to predict 2004 HCPM outputs.
- A small number of variables (e.g., some measure of the number of locations in the wire center, road feet) accounted for the majority of the regression equations' explanatory power.

SELECTION OF A FINAL REGRESSION EQUATION

The final regression equation chosen was one of the manual variable selection regressions. This regression equation's explanatory and predictive power rivaled that of the best "full variable" regression, but included only six variables – i.e., the final regression equation was simple, but not at an appreciable expense of predictive accuracy. Table 2 presents the estimated regression equation coefficients (calibrated on the full sample) and t-value statistics.[6]

Table 2. Selected Final Regression Equation

Variable	Definition	Coefficient	t-value
Intercept		7.08	87.08
LN_NearestCOft	Distance to nearest CO	0.02	7.40
LN_TOTAL_LOCS	Households+business locations in CLLI	-0.15	-8.62
LN_CLLIRoadft	Total road feet in CLLI	0.22	50.17
LN_TOTAL_LOCS_SQ	Households+business locations in CLLI, squared	0.06	51.23
LN_BizLocs_SQ	Business locations in CLLI, squared	-0.01	-20.09
LN_LocDensity	(Households+business locations in CLLI)/CLLI area	-0.07	-19.69

[6] The regression shown in Table 2 is a "log-linear" regression. To yield estimates of Total Cost, the predicted values of log cost must be converted back to dollars. The statistics shown in Table 1 are calculated after this retransformation.

The explanatory power of the regression was evaluated by comparing estimated costs produced by the regression to actual HCPM results for the holdout sample and the full sample. The adjusted r^2 value in both instances was 0.91.

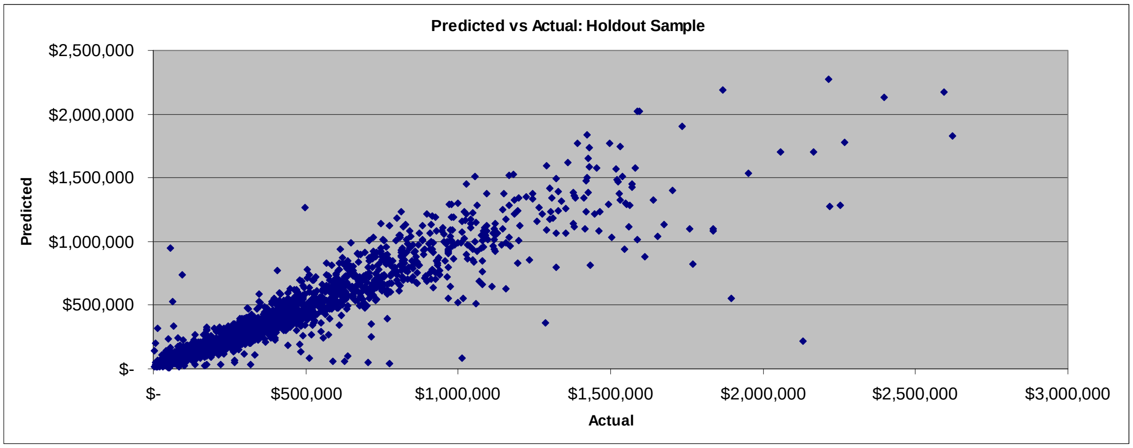
The final regression equation's ability to rank wire centers by their predicted cost is shown in Table 3 for the holdout sample. After sorting the wire centers from highest to lowest predicted cost and dividing into deciles, it is evident that the final regression exhibits a strong tendency to place wire centers in the correct deciles. That is, wire centers with the highest predicted cost are placed in the same decile as wire centers that actually have the highest actual cost as determined by the 2004 HCPM outputs.

Table 3. Final Regression Decile Performance in Holdout Sample

	Total Cost	
Decile	**Average Predicted By the Final Regression**	**Actual Average Determined by the HCPM**
1	$ 1,072,073	$ 1,055,807
2	$ 601,165	$ 616,122
3	$ 393,113	$ 411,115
4	$ 267,808	$ 289,754
5	$ 188,417	$ 204,614
6	$ 142,531	$ 145,920
7	$ 108,406	$ 115,072
8	$ 83,513	$ 88,711
9	$ 62,129	$ 67,263
10	$ 39,584	$ 45,247
Average	$ 295,731	$ 303,823

A visual of the final regression equation's performance is shown in Chart 1. Chart 1 presents the predicted and actual Total Cost for the holdout sample by wire center.

Chart 1. Final Regression Prediction Plot in Holdout Sample



ESTIMATING HCPM COSTS WITH UPDATED INPUTS

To estimate HCPM costs based on current conditions, the following input data sets were used to derive new driver variables: Tiger 2009 roads, current wire center boundaries (TANA 2010), Geolytics residential counts (2009), and GeoResults business counts by Census Block. These input data sets, constructed in the same manner as the data previously used for the regression equation development, represent a more recent vintage of driver variables for each wire center. When the updated driver variables are processed through the equation described above, the results of the equation yield a forecast of a Total Cost estimate for each wire center.

Once the Total Cost estimate for each wire center is developed, the next step is to develop a unitized cost per wire center. This figure can be compared to a benchmark cost per unit to determine whether a particular wire center is eligible for universal service support, and how much support each wire center would be eligible at any given benchmark level.

CONCLUSION

It is feasible to develop a simple regression equation that can be used to produce accurate estimates of HCPM outputs for all wire centers served by price cap carriers. Applying an updated set of input demographics, in conjunction with this statistical regression, makes it possible to estimate Total Cost in all such wire centers in light of current geographic and demographic conditions.